DOCUMENT RESUME

ED 225 479 HE 015 839

AUTHOR Bloom, Allan M.; Winstead, Wayland H.
TITLE Not Your Average Data Base. SAIR Conference Paper.
PUB DATE Oct 82
NOTE 20p.; Paper presented at the Annual Conference of the Southern Association for Institutional Research (Birmingham, AL, October 28-29, 1982).
PUB TYPE Reports - Descriptive (141) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computer Oriented Programs; *Databases; Higher Education; *Information Dissemination; Information Storage; Institutional Research; *Management Information Systems; Reports; State Universities; *Student Characteristics
IDENTIFIERS *SAIR Conference; *Virginia Polytechnic Inst and State Univ

ABSTRACT

The development of an unusual database at Virginia Polytechnic Institute and State University (Virginia Tech) is described. The Student Census-date Report File (STUCENFL) was designed to meet both internal and external data recipients' needs for student-related information. Attention is directed to needs for the database, underlying design concepts, the development process, and benefits. STUCENFL yields reports and ad hoc analyses that contain valid and internally-consistent information, and that cross-total to related reports and studies. The data items of STUCENFL are a subset of the student, academic, and timetable portions of the Student Data Base. One STUCENFL record is extracted for each student, containing demographic, diploma, admissions, academic history, and current course registration data. The data are extensively edited and updated to achieve valid, consistent, and reliable information. Details of the system, including data elements (BSR and USDS), derived report files, and records for six student populations are presented, along with a glossary of terms. Information for users of STUCENFL as a data source for automated systems is also included. (SW)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

SAIR

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED225479

NOT YOUR AVERAGE DATA BASE

Allan M. Bloom
Wayland H. Winstead

Management and Planning Analysis
Virginia Polytechnic Institute and State University
Blacksburg, Virginia 24061
(703) 961-6994

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

Abstract

Your average institutional data base often seems living proof of Sturgeon's Law: "95% of everything is crap." This is hyperbolic, but many data items of interest to Institutional Research are not critical to the operating divisions of the institution having authority over the data base. Operating divisions generally give those data secondary attention. Furthermore, even data items critical to the operating divisions' operations are subject to error, omission, and untimeliness. Critical data errors are corrected, but the corrections are not always prior to the census date at which the data base is "frozen" for reporting purposes. On-line information systems, and distributed data processing exacerbate the problem. Reports drawn directly from the data base can disagree if executed as little as five minutes apart. Even reports drawn from a "freeze" cross-total only accidentally, due to differences in program handling of invalid and inconsistent data.

This paper presents Virginia Tech's experience in developing a not-so-average institutional data base designed to meet both internal and external data recipients' needs for student related information. Our not-your-average data base yields reports and ad hoc analyses which contain valid and internally consistent information, and which cross-total to related reports and studies. Further, it does so with less human resource expenditure in data and report preparation.

The conditions which led to its development, the concepts underlying its design, the organization of the development effort, and the benefits derived from its creation are presented in detail.

HE 015 839

Paper presented at the Southern Association for Institutional Research Conference, October 27-29, 1982, Birmingham, Alabama

SOUTHERN ASSOCIATION FOR INSTITUTIONAL RESEARCH

This paper was presented at the 1982 Annual Conference of the Southern Association for Institutional Research held in Birmingham, Alabama, October 1982. It was reviewed by the SAIR Publications Committee and was judged to be of high quality and of interest to others concerned with the research in higher education. This paper has therefore been selected to be included in the ERIC collection of Conference Papers.

Gerald W. McLaughlin
President, SAIR

STUCENFL Users Guide

Executive Overview

The Student Census-date Report FiLe (STUCENFL, pronounced stoo-sen'-full) was developed by Institutional Research as a university resource to allow consistent, timely, and reliable reporting of student-related data. STUCENFL is a quarterly magnetic tape file based on the contents of the Student Data Base as of the On Campus "census date" each term, supplemented by Off Campus registrations as of that program's (later) census date. The data items of STUCENFL are a selected subset of the "student," "academic," and "timetable" portions of the Student Data Base. One STUCENFL record is extracted for each student, containing demographic, diploma, admissions, academic history, and current course registration data. There are really two files, the data base extract (called STUCENEX) and the STUCENFL "report file." STUCENFL is distinguished from STUCENEX in two important ways. First, students of little or no interest for internal or external reporting purposes are purged from STUCENEX prior to the creation of STUCENFL. Second, STUCENFL's data are subjected to extensive editing and updating to make those data valid, consistent, and reliable.

STUCENFL was developed to address a growing problem. As information from the Student Data Base was disseminated to an ever expanding constituency in ever expanding ways, it became more difficult to coordinate the University's reported data. Different offices provided different reports from different data sources at different times, often with different definitions, always with at least slightly different ways of handling non-perfect data. External constituencies in particular insist that reports cross-check, whether or not such is appropriate in any particular case. To avoid the impression that the university could not "give the same answer to the same question," we needed a single, internally consistent, well-defined data source for student-related reporting.

The Student Data Base *per se*, while perfectly functional, is an inappropriate source for this class of reporting and analysis. Many data items of interest to Institutional Research and to our constituencies have nothing to do with the functionality of the data base. A student can be admitted, registered, graded, and graduated utterly independent of the coding of his or her ethnic background. Further, even though non-functional data items are remarkably "clean," the number of data items that we keep relative to our students is huge. If the Student Data Base had Ivory Snow's legendary 99.44% purity, it would have over 100,000 erroneous data items. There are nowhere near that many. However, it is a certainty that a simple query to find the oldest and the youngest student on campus will yield one born in 1900 and one who will be born in 1999. Addressing those data errors that lie "in the statistical fringe" is what STUCENFL is all about.

There are three major elements of the STUCENFL system: (1) extraction of the desired data from the Student Data Base, (2) extensive editing of

those data, and (3) reporting those student data to our various internal and external constituencies. Selecting and extracting key data items was a major project in itself. However, the core of the STUCENFL system is the editing and updating of those data, the most thorough "cleanup" of student data in the history of the institution. Due to the large number of errors and to the limited resources available to investigate and correct them, most of the "cleanups" applied to STUCENFL are general-case assumptions. The conceptual approach underlying those assumptions is extensively discussed later, and it is a striking feature of the system.

The third part of the project is reporting. This involves creating report files from STUCENFL, rather than from the data base. Creating report files in this new manner is a great improvement. They are now extracted from a single data source containing critical reporting data that are consistent and valid. These derived report files can be used with existing programs in offices around the campus, requiring that no new reporting software be developed. Further, much existing software can be scrapped. The single source permits implementation of the state's Uniform Student Data System. Its report programs, nine to date, relieve Virginia Tech of the program design and maintenance responsibilities for those external reports.

Editing and updating of the data prior to production of the report files saves time and money. It also yields better information products. STUCENFL processing provides more extensive error detection and correction than have been previously available, and the automatic correction of invalid and/or inconsistent data assures that any corrections are performed consistently for all student records within a reporting period, and from one reporting period to the next.

The development of efficient, highly automated procedures for cleaning up a large report file is not a trivial task. The sound design of both the overall processing and the individual computer programs employed in each step of the processing requires a clear conceptual view of both what is possible and what is desirable.

Underlying Design Concepts

The automated "clean-up" of STUCENFL is the heart of the entire system. The large number of records processed, and the large number of errors, preclude manual verification of the data. Further, extensive human intervention would reduce reliability in the detection of error and in the appropriate corrections. This section describes the conceptual basis for the automated error detection and correction process. In the following discussion, all terms HIGHLIGHTED are described in the appended "Glossary."

The concepts of "validity" and "consistency" form the foundation for the automatic processing of STUCENFL. A VALID data element contains a value that belongs to the set of allowable values for that element. For example, the set of allowable Ethnic Background Codes consists of the num-

bers "1" through "7". The value "6" is a valid ethnic code since it belongs to the allowable set, and the value "9" is not valid. Symbolically, if "M" is the set of allowable entries for a datum, and if "x" is the value of the datum in a given record, then "x" is a valid entry if and only if { x | M }.

Validity alone is insufficient. Several data items may be individually valid and still be invalid when viewed as a whole. This is the concept of "consistency." By CONSISTENT data, we mean that the valid values contained in the fields for two or more data elements are allowable combinations. For example, "Virginia" and "North Carolina" are valid permanent home residence entries, and "In State" and "Out of State" are valid tuition status entries. Of the four possible combinations of those entries, only two ("Virginia" - "In State" and "North Carolina" - "Out of State") are consistent with each other. Symbolically, if "M" and "N" are sets of individually valid entries in two related data fields, then the consistency of the entries may be defined by a truth table formed by the Cartesian product "M" X "N". If the "mth" value in "M" is consistent with the "nth" entry in "N", then the matrix entry "m,n" is "1", else it is "0".

One may extend this representation of consistency to multi-way tests, involving more than two sets, by letting "MN" denote the set of consistent combinations of the values of "M" and "N", and by letting "L" denote the set of valid entries in a third, related field. The consistency of value "1" would be determined by an entry of "1" for element "mn,1" in the truth table formed by "MN" and "L". Any number "x" of related data values can be tested for internal consistency by "x-1" sequentially formed truth tables.

STUCENFL processing keys to the validity and consistency of data. This was implemented by defining sets of rules (EDIT criteria) for a computer program to apply to the elements of the STUCENEX file. A comprehensive list of violations of those rules is reported to the appropriate university administrative offices for review and possible correction. However, we must live with the realities of census-date reporting and the limited resources available for error investigation and correction. STUCENFL processing can not be dependent on actual individual-datum error correction by other offices, and it is not so dependent. Whenever an invalid or inconsistent condition occurs, another set of rules is applied to correct the situation.

Having an invalid or an inconsistent data value is a situation that can be corrected by defining a set of MASS UPDATE criteria. A computer program can replace an invalid or inconsistent datum with one obeying the applicable edit criteria. That is not necessarily the same as correcting the error, since a valid and consistent data value is not necessarily ACCURATE. Accurate data must also map the attributes of the real world truthfully, and nothing short of a comprehensive data AUDIT can ensure that the computer representation is a mirror of the real world. However, a reasonable and defensible mass-update criterion can yield a replacement value that is superior to the original and that is in time to prepare reports and analyses containing valid and consistent data summaries. Such general case assumptions are applied to a small percentage of the data. Most mass update processing is expended in converting known obsolete codes.

The mass update of STUCENFL is distinguished by being entirely automatic, performed without human intervention. If a datum is found to contain an invalid entry, then the invalid entry is automatically replaced by a valid DEFAULT value. The determination of the default value is based principally on experience with the type of errors which have occurred historically. Where experience fails, a "most common" value is assigned. After ensuring individual data element validity, processing continues with the consistency of related data.

When inconsistencies occur, resolution follows a convention based upon the sequential algorithm used in multi-way editing. Historical experience has shown that some data elements are more reliable, in general, than others. Where no such confidence precedence is known, it is assigned via a conservative policy — one that does not enhance any critical counts. Inconsistencies are resolved by bringing the lower-confidence datum into line with the higher-confidence one. In multi-way inconsistencies, processing is cascaded. After the two highest-confidence data values are made consistent, the third is aligned with the first two.

The automatic updating of data elements to achieve consistency is constrained by the formal logical structure used in processing the file, and conditions arise which preclude the automatic correction of some inconsistent entries. In those cases no automatic updates are performed, and resolution of the inconsistency is left to human intervention. That may be in the form of adding a mass-update algorithm to handle the problem or as an INDIVIDUAL UPDATE.

Individual updating, as the term implies, is distinguished by the human intervention required to identify and/or resolve particular residual and potential errors in individual records. If a class of errors is detected after mass updating, an algorithm is developed to automatically treat the problem in the next processing cycle. By continuing to refine the mass-update part of the process, very few individual anomalies need human attention on a regular basis. A further refinement is the historical individual update. After errors are ferreted out via the appropriate administrative offices, and after the individual correction is applied in the current processing cycle, the error and its known correction are added as a part of the next mass update. If an individual error recurs, it is trapped and automatically corrected. If the error had been corrected between cycles, the historical update -- automatically noted as superfluous -- is removed.

To summarize, STUCENFL processing is based on emplacing the most rigorous possible automatic edit and update mechanisms to assure that the data in the file are valid and consistent with a minimum of human resource expenditure. No comprehensive audit for accuracy is performed, that task being reserved for the administrative offices controlling the data base. Within the constraints which occur in the real world, it is not reasonable to expect that all of the data in the student data base, or in STUCENFL, will be accurate at any given moment. Census-date reporting must needs be

based on data as of a moment in time. Given this realistic expectation, the concepts of valid and consistent data which are reliable over time represent what we believe to be the most appropriate basis for the design and processing of a "clean" report file.

System Summary

The STUCENEX initial extract contains critical report data for everyone on the Student Data Base (SDB), over 100,000 records. The STUCENFL report file contains a smaller subset of the population, some 70,000 records of active students. Two subsets of the students represented in SDB are purged in the process of creating STUCENFL. The most extensive subset of purged records describe students who have not attended, or graduated from, the university within the last five years (25 academic terms). In addition, records describing applicants for admission in prior (non-current) academic years are also purged. These records are purged since they are of little or no interest in internal or external reporting. Further, the reduction of approximately 30,000 records greatly expedites the extensive processing inherent in STUCENFL, thus reducing both effort and cost.

STUCENFL contains records for four major subsets of the remaining student population, including students enrolled in both On- and Off-Campus credit programs, past students who last attended, or graduated from, the institution within the previous five years, and never-attended applicants whose anticipated entry is within the current admissions year. The file is based on the contents of the SDB as of the On-Campus "census date" each term (10 calendar days after check-in for the Academic Year terms, five days afterward for each Summer Session). The file is not complete until the Off-Campus census date (two calendar weeks after the On-Campus census date during the Academic Year, one week during the Summer Sessions), when enrollment data for Off-Campus students are added to STUCENFL. There are six STUCENFL report files created per year, one for each academic term, plus a "combined summer" file.

While STUCENFL is an extract of the Student Data Base (SDB), the values of its data items are not necessarily the same as their corresponding SDB entries. Any "live" data base contains invalid data at any given time, and SDB is no exception. During the extensive edit and update processing of the file, which has been previously described, these invalid data are updated so that STUCENFL contains both valid and consistent data. Further, there are certain "reporting conventions" that require viewing certain perfectly good SDB data in a different light. For simplicity and consistency, those conventions are also built into the file. Since both of these elements are incorporated in the "clean" file, there is no longer the possibility that reports will vary due to to application of either differing assumptions or different conventions.

Using STUCENFL

This section is aimed at those who use STUCENFL as a data source for electronic data processing. It contains useful information not found in, and to be used in conjunction with, three other documents (available from Institutional Research, X6994):

1. STUCENFL Data Element Dictionary: a description of each data item in the file, its definition and format, edit criteria, and mass-update processing of inconsistent or invalid data.
2. INSRES CODES: a table of information on Virginia Tech departments, majors, courses, and degree programs.
3. INSRES XLATES: a table of Virginia Tech coded data lists (state abbreviations, residence home codes, Virginia ZIP codes, etc) and preferred translations for reporting.

All four documents should be used as a set, and they should be sufficient to permit instructing the computer in the selection and processing of STUCENFL records appropriate to a given task.

STUCENFL is an 877-character record whose PL/1 record description is member "STUCENFL" of CMS file PLSOURCE MACLIB on user-ID "OIRSOFT" (read password "BLOOM"). The extract file (STUCENEX) contains about 100,000 records. The report file (STUCENFL) tape contains some 70,000 records, and the corresponding active-student subset on disk has about 30,000 records. Each file is sequenced by ascending social security number (record positions 12-20).

There are ten (10) initial data base extract files created during the year, two per term. The first captures the entire student data base as of the ON Campus census date. The second, on the same physical volume as the first, captures only students with OFF Campus course registrations as of the OFF Campus census date. These initial extract files are identified by having "STUCENEX" as a part of their data set names. The six (6) report files; with ON and OFF Campus enrollment data merged and pre-processed for reporting, are identified by the "STUCENFL" in their data set names. The data set names are

STUCENEX.ttyy	The full-data-base extract file, taken at the On Campus census date (tt= FA,WI,SP,FS,SS).
STUCENEX.ttyy.OFF	The Off-Campus extract file (tt= FA,WI,SP,FS,SS) Label 2 of the STUCENEX.ttyy tape.

STUCENFL Data Sets (Continued):

A713F7.STUCENFL. CYCLE (x)	The general update-cycle file. Cycled through mass and individual updating until "clean." Three-generation tape file.
A713F7.STUCENFL. OFFCYCLE (x)	The Off-Campus update-cycle file. Cycled through mass and individual updating until "clean." Two-generation disk file on INSRES pack.
STUCENFL.ttyy	The combined cleaned up student census file for a term (tt=FA,WI,SP,FS,SS,SU), a 6250 bpi tape file. Results from a merger of the CYCLE(0) and OFFCYCLE(0) files for a regular term, or from merging "FSyy" and "SSyy" for combined summer.
A713F7.STUCENFL. CURRENT	The current-term active-student file cataloged on general-use disk USER05. Has only records with current-term credits greater than zero. Updated from CYCLE(0) after it is clean, and from STUCENFL.ttyy after its creation.

Specific tape volume information is maintained by Institutional Research, X6994. It may also be found in the DATASETS TAPE file contained on CMS user-ID "OIRSOFT." The DATASETS TAPE file contains the following information for each Institutional Research tape file:

Col 01 - 05 Tape volume serial number
Col 06 Tape density code: " " = 1600 bpi, "+" = 6250 bpi, "-" = 800 bpi.
Col 07 - 08 Tape label number (file number within volume).
Col 10 - 13 General data category: "SCF " = STUCENFL-related files.
Col 15 - 50 File data set name (DSN), plus optional comments.
Col 51 - 56 File creation date, in "YYMMDD" format.
Col 58 - 60 Person responsible for data file, initials.
Col 61 - 66 Record count, number of logical records in the file.
Col 67 - 68 Recording format (RECFM).
Col 69 - 72 Logical record length (LRECL).
Col 73 - 79 Block (physical record) length (BLKSIZE).

STUCENFL consists of records for six populations of students and applicants. Those groups are as follows:

ON	Currently enrolled ON-Campus students
OFF	Currently enrolled OFF-Campus students
ONCE	Students who have ONCE attended the University, within the preceding five years in the report file.
NEVER	Applicants who have NEVER attended the University, with anticipated entry within the current calendar year in the report file.
ON (+)	Current ON-Campus students registered for zero total credits and therefore not reportable as "enrolled."
OFF (+)	Current OFF-Campus students registered for zero total credits and therefore not reportable as "enrolled."

Due to reporting conventions, only students registered for credit are reported as "enrolled." We do, however, have a number of students each term registered for zero credits--principally Cooperative Education Program students in the "work" part of their "work/school" cycle and students who are auditing all their courses. For certain applications (student retention analysis, for example) it is useful to know that those students are "still with us," even though we may not report them as enrolled. Therefore STUCENFL permits identifying them.

All six populations are selected from CDB on the On-Campus census date. Data for the OFF's are neither complete nor usable for official University reports until after the Off-Campus census date enrollments are merged into STUCENFL. Data for the NEVER's accumulate during the calendar year and are considered complete as of the Fall Term On-Campus census date for admissions analysis reporting. The ONCE population consists of students who last attended the University or last graduated within the past five years.

Selection of a desired population is keyed to four (4) STUCENFL record fields:

TAPE_QYY	FILE_IDENTIFICATION.TAPE_QTR_YR (Current Term)
ON_QYY	STU31_DATA.LAST_ENROLLMENT_DATE (On Campus)
OFF_QYY	STU31_DATA.OFF_CAMPUS_LAST_ENROLLMENT
CREDIT_TOTAL	The total number of credits as of the census date.

The following "decision table" is used to select a desired population from STUCENFL:

Population	ON_QYY	OFF_QYY	CREDIT_TOTAL
ON	= TAPE_QYY	--	> 0
OFF	--	= TAPE_QYY	> 0
ONCE	< TAPE_QYY	< TAPE_QYY	---
NEVER	= 0	= 0	---
ON(+)	= TAPE_QYY	--	= 0
OFF(+)	--	= TAPE_QYY	= 0

Additionally, the data base extract program (STUCENEX, OIR120) can optionally select several different subsets of the Student Data Base for special needs. The processing options are

1. All students.
2. All On Campus students (STU31 Campus Location = '00').
3. All Off Campus students (STU31 Campus Location > '00').
4. All Currently enrolled students (course registrations > 0).
5. Currently enrolled On Campus students (at least one course registration with Course Location = '00').
6. Currently enrolled Off Campus students (at least one course registration with Course Location > '00').

The office routinely uses options "1" and "6" only. Other extracts may be performed on request. The requestor is responsible for the cost of any special extract and for any data clean-up to be performed.

While latest term registered and CREDIT_TOTAL will yield currently registered students, the CREDIT_TOTAL field will not necessarily yield the student's On Campus or Off Campus credits. It is the sum of ALL credits. Each quarter finds a handful of students who are registered for both On and Off Campus credits. To differentiate, the programmer must scan through the registration segments and look at the course location in each ("00" for On Campus). SCHE' guidelines state that a student with ANY On Campus credits is an On Campus student. For compatibility with USDS reports, therefore, one may use the CREDIT_TOTAL field, but the OFF Student definition would be modified to EXCLUDE any students with ON_QYY = TAPE_QYY.

Derived Report Files

For convenience and for meeting the needs of externally-provided reporting systems, several "derived" files are produced regularly from STUCENFL. They are as follows:

QSR — Quarterly Student Record tape, a file of long standing for internal and external reporting, one record per currently enrolled student, extracted for ON- and OFF-Campus enrollees each quarter, and for combined ON- and OFF-Campus enrollees on request. There are a few BSR data items that are not available from STUCENFL:

BSR_SS_STATUS	Selective Service Status	Pos 94
BSR_SS_xxxxx	Other Selective Service Info.	Pos 123-134
BSR_PARENT_NAME	Name of the student's parent.	Pos 210-230
BSR_PARENT_PHONE	Name of the student's parent.	Pos 210-230
BSR_STANDING	Class Standing in Dept & Coll.	Pos 418-459
BSR_ACH_xxxx	SAT Achievement Test Scores.	Pos 507-516

Further, BSR_SEX (Pos 84) contains only "sex." It does not indicate marital status as well as sex.

Program SCBSR is optioned to count only ON, only OFF, or ALL credits. Selecting only ON or only OFF Campus students from STUCENFL without using the option will mis-assign registrations for students taking both ON and OFF Campus courses in the same term. The number isn't large, but it can mess up an analysis.

SDM — Student Data Module tape, used as input to the Commonwealth's enrollment reporting system, one record per currently enrolled student per course registration, extracted for ON- and OFF-Campus enrollees each quarter. Using STUCENFL, the following data have been added to SDM:

RESIDENCY_CODE	"I" or "O" for In/Out-of-state	Pos 26
COURSE_LOCATION	"00" or 2-digit Consortium Code	Pos 27-28

USDS — Uniform Student Data System tape, used as input to the Commonwealth's standard package for producing the following external reports:

- R1 County or city of residence of students enrolled in Virginia institutions of higher education
- B2 Summary on-campus headcount enrollment by term
- B3 Summary off-campus headcount enrollment by term
- B5 Headcount enrollment age summary
- 2.1 OCR2300-2.1 Degrees and other formal awards conferred
- 2.1s Supplement to OCR2300-2.1

2.3 OE2300-2.3 Fall enrollment
2.3s Supplement to OE2300-2.3
2.8 OE2300-2.8 Residence and migration of college students

USDS is extracted each Fall (one record per currently enrolled ON and OFF-Campus student) and at end-Summer (one record for each graduate in the preceding July-June Fiscal Year). The fiscal-year file is drawn from a subset of STUCENFL selected by program SCDIPL. That program looks at each Tech Diploma (STU414) segment to see if at least one degree was awarded during the fiscal year specified via option card. In no other way can all diplomas awarded during a period of interest be selected.

The first two files (QSR and SDM) have been in use for quite some time, and the only significant change is their extraction from a single file whose critical reporting data are consistent and valid. Editing and updating are performed (on STUCENFL) prior to SDM and QSR production instead of to each file afterward. As a bonus, STUCENFL editing is more extensive than that previously performed on either file. Both files contain more valid and consistent data as a result.

USDS has no Virginia Tech antecedent, per se. It and the Commonwealth external report package, however, replace several files and programs previously used to generate many state and federal reports. University personnel need no longer bother with those files and programs, a significant decrease in resource expenditure. Further, with USDS those reports are forced to be consistent with each other.

The remainder of this document is divided into appendices giving detailed information on STUCENFL and its related processing.

Appendix A. Glossary of Terms Used in This Document.

Appendix B. USDS Data Elements Obtained from STUCENFL by Program SCUSDS

Appendix C. BSR Data Elements Obtained from STUCENFL by Program SCBSR

Glossary of Terms Used in This Document

ACCURATE A datum is accurate if it is valid and if it is true. For example, if a black female is coded "black female," the data are accurate.

AUDIT One audits a file to see whether critical data are accurate. An audit is performed on both invalid and valid data, whether consistent or not. Only an audit, a minute examination of evidence, can determine that a person coded "black female" is actually a black female.

CONSISTENT Data that are individually valid may be inappropriate when viewed in context. For example, "Virginia" and "North Carolina" are valid permanent home residence entries, and "In State" and "Out of State" are valid tuition status entries. Of the four possible combinations of those entries, only two ("Virginia" - "In State" and "North Carolina" - "Out of State") are consistent with each other. Individually valid data, taken together, must make sense to be consistent.

DEFAULT If a datum is invalid, mass updating assigns a reasonable, valid value. That assigned value is the default for that datum, either the most commonly occurring of the set of valid values or one that has has been shown most probable by experience.

EDIT One edits a file to see whether critical data conform to a set of rules defining data validity and consistency. The process is analogous to checking a document for spelling and grammar.

UPDATE One updates a file to change its data entries. This may arise from an audit or an edit which found inaccurate, inconsistent, or invalid data, or from a change in status. The term "posting" is used for initial data entry, a process not addressed in this document.

MASS updating is performed on a class of data records, say to convert Virginia residence codes for all out-of-state students to codes for other states. A "mass update" change to an invalid or inconsistent datum is derived from a basis of known valid data and a set of rules governing their relationships. INDIVIDUAL updating is applied to a specific record to correct a specific problem. In STUCENFL processing, individual updates are made only for those few cases wherein a general-case mass update cannot be described.

VALID A datum is valid if its content belongs to the set of allowable values, as prescribed by data base administration policy. Symbolically, if M is a set of allowable entries in a field, and x is an entry in the field, then x is a valid entry if and only if $\{x|M\}$.

USDS Data Elements Obtained from STUCENFL by Program SCUSDS

USDS Data Element	Processing in SCUSDS using STUCENFL
1. Institution FICE Code	VA Tech's FICE Code = "003754."
2. Reporting Year-Fiscal Format: y1y2	Converted from TAPE_QTR_YR: YR-1 || YR if QTR 0, 2 or 3 YR || YR+1 if QTR 4, 5, 6, or 1
3. Reporting Term	Converted from TAPE_QTR_YR: 01 if QTR 6 02 if QTR 1 03 if QTR 2 04 if QTR 3 05 if QTR 0
4. Student identifier	Student social security number.
5. Name of Student	Concatenated student name fields, last-first-middle-generation, with multiple blanks removed.
6. Sex of Student	Converted from STU21_DATA.SEX: 01 if SEX "M" 02 if SEX "F"
7. Race or Ethnic Category	STU21_DATA.ETHNIC_BACKGROUND, from 1 to 2 digits, except "06" if "7".
8. Degree #1 Program ID	Latest Tech diploma's IMS HEGIS Code translated to SCHEV NCES Code. Blank if no Tech diploma.
9. Degree #1 Level	Latest Tech degree abbrev translated to SCHEV degree level. Blank if no Tech diploma.
10. Degree #2 Program ID	Next-most-recent Tech degree awarded IMS HEGIS Code translated to SCHEV NCES Code, IF awarded in the same Fiscal Year as Degree #1, ELSE blank.
11. Degree #2 Level	Next-most-recent Tech degree abbrev, translated to SCHEV degree level. If Degree #2 is blank, this is blank.
12. Student Major	Translate STU31_DATA.HEGIS_CODE_MAJOR To SCHEV NCES code.
13. Student Level	Translate STU31_DATA.ACADEMIC_LEVEL_ CURRENT to SCHEV student level code.

USDS Data Elements Obtained From STUCENFL by Program SCUSDS

USDS Data Element	Processing in SCUSDS using STUCENFL
14. Second Student Major	Translate ACA27_DATA.DOUBLE_OR_SECOND_MAJOR.CURR_ABBREV to SCHEV NCES Code.
15. Birth Year of Student	From STU21_DATA.BIRTH_DAY.
16. Citizenship of Student	From STU21_DATA.ETHNIC_BACKGROUND: Ethnic 6 = 03 Non-resident alien Ethnic 7 = 02 Resident alien Else 01 U.S. citizen
17. State Residency of Student	Translate digits 1 and 2 of STU31_DATA RESIDENCE_HOME_CODE to USDS Code.
18. City-County Residency	Translate digits 3 thru 5 (county code) or digits 6 through 9 (city code) of STU31_DATA.RESIDENCE_HOME_CODE for VA residents to USDS code. Else "0904."
19. Enrollment Status of Student	A student is NEW if: STU21 FIRST_ENROLLMENT = TAPE_QTR_YR or ACA11 QUARTER_GRADE_SUMMARY term = 000. For Fall, FIRST_ENROLLMENT during the Summer is also NEW. A NEW TRANSFER student is NEW and has STU31_DATA.TRANSFER_STUDENT_FLAG = "Y". For other students, if the difference TAPE_QTR_YR) and last previous enrollment (ACA11 QUARTER_GRADE_SUMMARY term) is 1 (less than 4 terms in the Fall) the student is CONTINUING, else READMITTED.
20. Total On-Campus Credit Hrs	The sum of course credit hours (ACA22_DATA.CREDIT.HOURS) on campus: TIM11_DATA.CAMPUS-LOCATION-NUM = 00.
21. Total Off-Campus Credit Hrs	The sum of course credit hours (ACA22_DATA.CREDIT.HOURS) off campus: TIM11_DATA.CAMPUS-LOCATION-NUM > 00.

BSR Data Elements Obtained from STUCENFL by Program SCBSR

BSR Data Element	Processing in SCBSR using STUCENFL
1. BSR_STATUS	Always 'A' for admitted status
2. BSR_SOC_SEC_NO	STU11_DATA.SOCSECNUM
3. BSR_HONOR_STUDENT	STU31_DATA.HONORS_PROGRAM_FLAG
4. BSR_TYY_FIRST_ATTEND	STU21_DATA.FIRST_ENROLLMENT_DATE
5. BSR_TYY_LAST_ATTEND	ACA11DATA.QUARTER_GRADE_SUMMARY
6. BSR_TYY_GRADUATION	STU31_DATA.GRAD_DATE_ACTUAL
7. BSR_ELIG_TO_RETURN_DATE	RETURN_DATE in ACA24_DATA.ACADEMIC_DROP_SECTION or OTHER_LEAVE_SECTION, whichever is more recent
8. BSR_LAST_NAME,FIRST_NAME, MIDDLE_NAME,SUFFIX	From STU2223_DATA
9. BSR_BIRTH_DATE	Translated from STU21_DATA.BIRTH_DAY (MMDDYY) into YYMMDD format
10. BSR_BIRTH_PLACE	Last two digits from STU21_DATA. BIRTH_PLACE.COUNTRY_CODE if student was born in U.S., else translated from COUNTRY_CODE into a two digit code via XLATES, else (if unable to translate) the last two digits from COUNTRY_CODE
11. BSR_SEX	If STU21_DATA.SEX is male, then '1' (single male) else '3' (single female)
12. BSR_REG_CODE,CURR_ABB	Translated from STU31_DATA.HEGIS_CODE_MAJOR using INSRES CODES table
13. BSR_TUITION	If STU31_DATA.RESIDENCY_CODE is 'I', then '2' (in-state) 'O', then '1' (out-of-state)
14. BSR_LOCATION	If (STU31_DATA.CADET_FLAG, STU21_DATA.SEX, STU31_DATA.DORM_STUDENT_FLAG) are (1,M,N) then 1 (1,M,Y) then 2 (2,M,N) then 4 (2,M,Y) then 3 (1,F,Y) then 5 (1,F,N) then 6 (2,F,Y) then 7 (2,F,N) then 8

BSR Data Elements Obtained from STUCENFL by Program SCBSR

BSR Data Element	Processing in SCBSR using STUCENFL
15. BSR_TRANSFER	If STU31_DATA.TRANSFER_STUDENT_FLAG is 'Y', then translate SCHOOL_CODE (using college code table) of the first STU414_DATA segment into a two-digit code (if unable to translate, then 'Z1'; if no STU414 segment exists, then blanks). '00' for non-transfers
16. BSR_SUPERIOR_STUDENT	STU5111_DATA.SUPERIOR_STUDENT_CODE for undergrads, else blank
17. BSR_DECEASED_FLAG	STU21_DATA.DECEASED_STUDENT_FLAG
18. BSR_HIGH_SCHOOL	STU411_DATA.SCHOOL_CODE
19. BSR_STUDENT_LEVEL_1,LEVEL_2	First and second digits of STU31_DATA. ACADEMIC_LEVEL_CURRENT, respectively
20. BSR_CAMPUS_CODE	STU31_DATA.CAMPUS_LOCATION_NUM
21. BSR_OFF_CAMPUS_LAST_ENROLL	STU31_DATA.OFF_CAMPUS_LAST_ENROLLMENT
22. BSR_RESIGN_DATE	From ACA24_DATA.ACADEMIC_DROP_SECTION. OTHER_LEAVE_SECTION.RESIGNATION_DATE: '1yy' -> 'yy0930' '2yy' -> 'yy0131' '3yy' -> 'yy0331' '4yy' -> 'yy0630' '5yy' -> 'yy0730' else blanks
23. BSR_RESIDENCE_CODE	STU31_DATA.RESIDENCE_HOME_CODE
24. BSR_DORM_NUM,ROOM_NUM	From ADDRESS_DORM.FACILITY_NUM and ROOM_NUM. If STU31_DATA's DORM_STUDENT_FLAG is 'Y', else blanks
25. BSR_LOCAL_ADDRESS,CITY, STATE,ZIP,PHONE,ADDR_2	From STU32_DATA.LOCAL_ADDRESS
26. BSR_PARENT_ADDR1,ADDR2, CITY,STATE,ZIP	From STU32_DATA.PERMANENT_ADDRESS. If ADDRESS_2 is blank and ADDRESS_1 is not blank, then ADDR1 and ADDR2 are exchanged. If not a U.S resident (STATE_ABBREV = '99'), the last word (if any) of PERMANENT_ADDRESS.CITY is used for STATE, and the beginning words are used for CITY.

BSR Data Elements Obtained from STUCENFL by Program SCBSR

BSR Data Element	Processing in SCBSR using STUCENFL
27. BSR_QTR_GRADES.	From ACA11_DATA.QUARTER_GRADE_SUMMARY, except that BSR_HOURS_ATTEMPTED is for the current academic quarter: If OPTION "ON ", sum of On Camp Hours OPTION "OFF", sum of Off Camp Hours OPTION "ALL", move CREDIT_TOTAL.
28. BSR_OVERALL_GRADES.	From ACA11_DATA.OVERALL_GRADE_SUMMARY
29. BSR_RACE	STU21_DATA.ETHNIC_BACKGROUND
30. BSR_ADM_DATE	If STU21_DATA.FIRST_ENROLLMENT_DATE is '1yy', then '0901yy' '2yy', then '0101yy' '3yy', then '0301yy' '4yy', then '0601yy' '5yy', then '0701yy' anything else, then zeros
31. BSR_GRAD_YR	STU411_DATA.GRAD_DATE_ACTUAL
32. BSR_CTR_CITZ	STU21_DATA.CTRY_CITIZENSHIP
33. BSR_CLASS_RANK	STU411_DATA.CLASS_RANK
34. BSR_CLASS_SIZE	STU411_DATA.CLASS_SIZE
35. BSR_MATH	STU5112_DATA.SCORE_SAT_MATH
36. BSR_VERBAL	STU5112_DATA.SCORE_SAT_VERBAL